NCBI Magic-BLAST

An efficient command line aligner for next generation sequence reads

https://ftp.ncbi.nlm.nih.gov/blast/executables/magicblast/LATEST

National Center for Biotechnology Information • National Library of Medicine • National Institutes of Health • Department of Health and Human Services

Overview

Magic-BLAST (1) is a tool for mapping large next-generation RNA or DNA sequencing runs to a genome assembly or transcriptome. Unlike other BLAST nucleotide search programs, such as BLASTN or Megablast, Magic-BLAST produces spliced alignments and optimizes alignment scores for paired reads. In addition to the BLAST code base, Magic-BLAST also incorporates additional ideas developed in the NCBI Magic pipeline, in particular hit extensions by local walk and jump. This approach is faster and more memory efficient than then the standard Smith-Waterman extension procedure. It directly accesses reads stored in the NCBI Sequence Read Archive (SRA), without the need to download the data beforehand, and reports alignments in the Sequence Alignment/Map (SAM) format and a tabular format, similar to BLAST tabular output.

Access and Features

Magic-BLAST packages for common platforms are freely available from NCBI FTP site (ftp.ncbi.nlm.nih.gov/blast/executables/magicblast/). Magic-BLAST is versatile in that it can take next generation (next-gen) sequencing data in various formats as input query. For those public datasets already deposited in the Sequence Read Archive (SRA) database, the prefetch function built-in can retrieve the data remotely from NCBI. By default, magic-BLAST presents alignment results in the widely used SAM format that can be used with samtools (2) for further processing to generate indexed bam file. Variant calls (vcf) can also be generated by using bcftools (2) with the BAM file. A video tutorial on magic-BLAST is available at: https://youtu.be/LrOHT73czZw.

Example Use Cases

1. Mapping next-gen reads to the Ebola genome and viewing results in the graphical sequence viewer

Ebola virus is a serious infectious agent that cause hemorrhagic fever in humans. The Reference Sequence (RefSeq) genomic entry for Ebola virus, NC_002549.1, represents the Zaire isolate from the outbreak in 1995. In the recent outbreak between 2014 - 2016 in West Africa, a lot more sequence data were collected using next-gen sequencing technology. Comparing these newly derived sequence data against the genome from a previous outbreak can yield valuable epidemiological and evolutionary information that could help further our understanding of the biology of this important pathogen. A sequencing run from one such dataset is SRR1553459 isolated in 2014 from Sierra Leone.

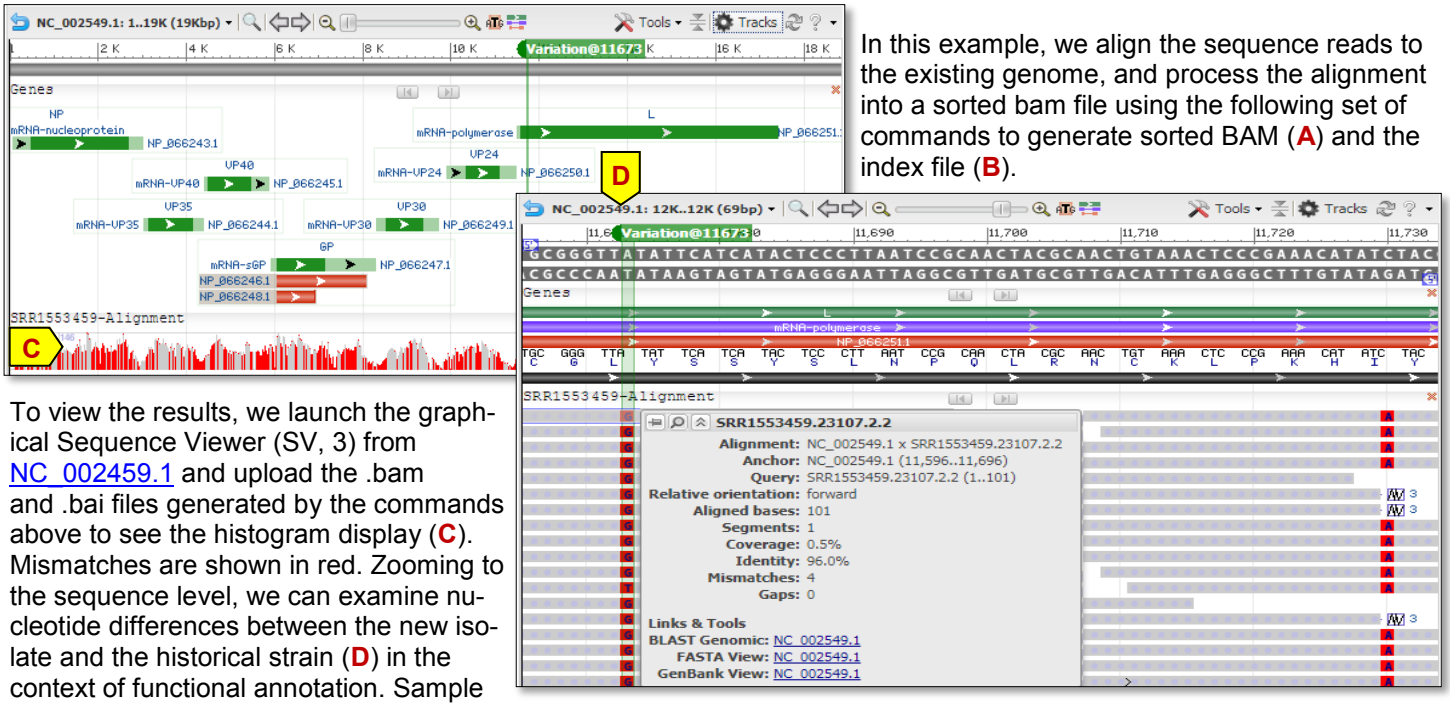
```
$ magicblast -sra SRR1553459 -subject NC_002549.acc -paired -splice F | \
samtools view -bS | \
samtools sort  -O BAM -o  SRR1553459-NC.bam

$ samtools index SRR1553459-NC.bam SRR1553459-NC.bai
```

A

B

In this example, we align the sequence reads to the existing genome, and process the alignment into a sorted bam file using the following set of commands to generate sorted BAM (**A**) and the index file (**B**).



To view the results, we launch the graphical Sequence Viewer (SV, 3) from NC_002459.1 and upload the .bam and .bai files generated by the commands above to see the histogram display (**C**). Mismatches are shown in red. Zooming to the sequence level, we can examine nucleotide differences between the new isolate and the historical strain (**D**) in the context of functional annotation. Sample files, SRR1553459-NC.bam and SRR1553459-NC.bai, are available at: https://go.usa.gov/xPFkU.

Contact: blast-help@ncbi.nlm.nih.gov

Example Use Cases (cont.)

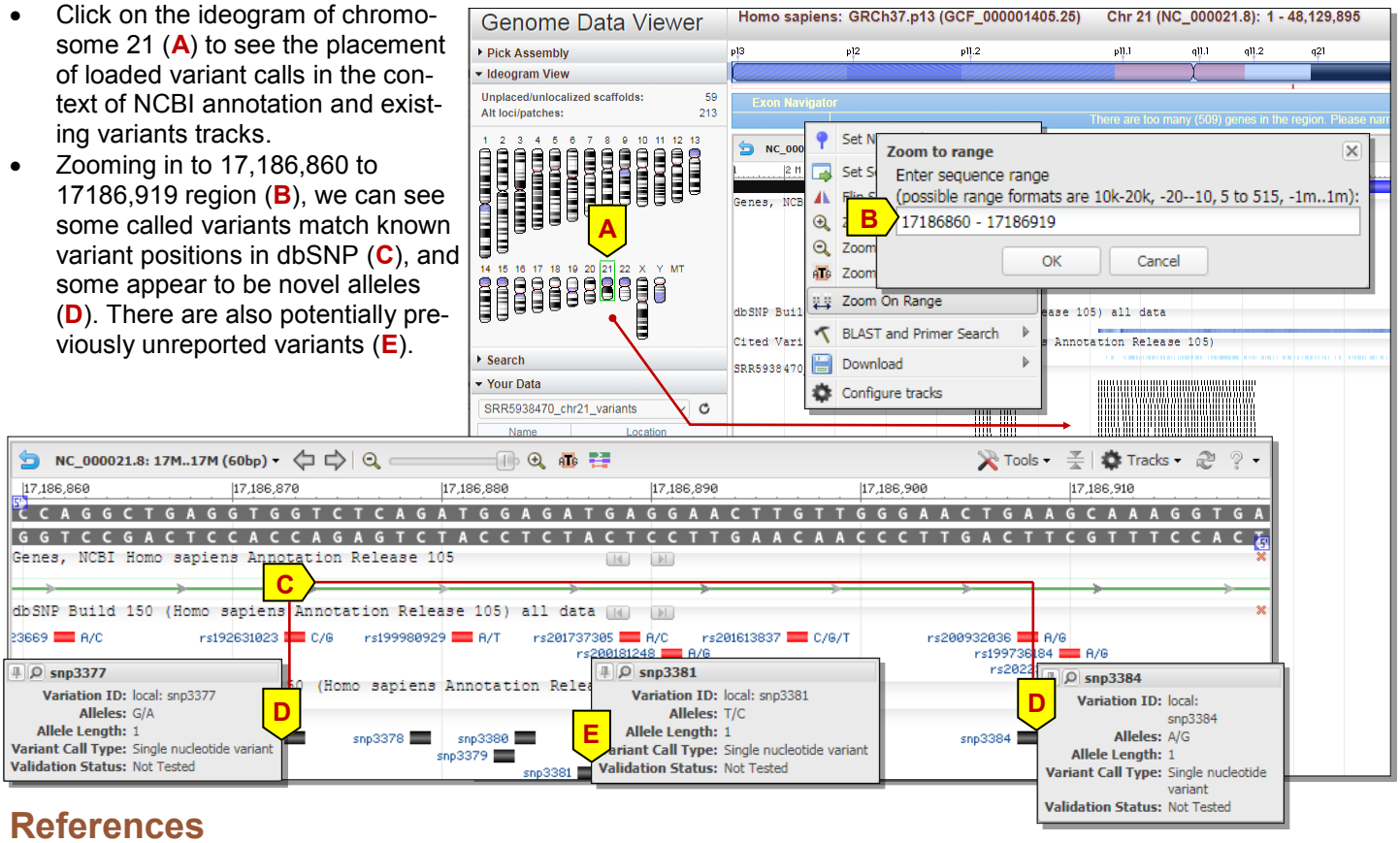
2. Aligning reads from brain control sample for variant calls and displaying the results in Genome Data Viewer

Datasets from SRA database can serve as useful control set or as additional samples for research. One read from an example dataset is SRR5938470 from a human control brain sample. We can use magic-BLAST to map the reads to human genome assembly. Using the samtools and bcftools, we can transform the alignment result in SAM format to BAM format, and generate variant calls from the BAM file. We can upload and view in the Genome Data Viewer (GDV, 4).

We are using GRCh37.p13, since many of the human clinical studies are still mapped to the previous assembly. For simplicity, we will only use chromosome 21 in this example.

Steps used to generate the intermediate results, upload the final vcf, and activate the display are given below.

- Download the FASTA of GRCh37.p13 chromosome 21
  $ efetch -db nuccore -id NC_000021.10 -format fasta > chr21.fa
- Run magic-BLAST to map the reads
  $ magicblast -sra SRR5938470 -subject chr21.fa -splice T -out SRR5938470_chr21.sam
- Generate sorted SAM file
  samtools view -bS SRR5938470_chr21.sam | samtools sort -O BAM -o SRR5938470_chr21.bam
- Create variant calls from the SAM
  $ samtools mpileup -uBI -f chr21.fa SRR5938470_chr21.bam | bcftools view - > tmp.vcf
- The resulting VCF file, edited to remove extra fields (steps not shown), is available at:
  ftp.ncbi.nlm.nih.gov/pub/factsheets/chr21.vcf
- Go to the GDV landing page (www.ncbi.nlm.nih.gov/genome/gdv/), select GRCh37.p13 from the Assembly menu
  www.ncbi.nlm.nih.gov/genome/gdv/browser/?acc=GCF_000001405.25&context=genome
- Use “Your Data >> “+” >> “Add URL” to activate the upload feature. Paste in the above URL, name the track if desired, and click “Upload” to load the above data file
- Click on the ideogram of chromosome 21 (**A**) to see the placement of loaded variant calls in the context of NCBI annotation and existing variants tracks.
- Zooming in to 17,186,860 to 17186,919 region (**B**), we can see some called variants match known variant positions in dbSNP (**C**), and some appear to be novel alleles (**D**). There are also potentially previously unreported variants (**E**).



References

1. Magic-BLAST git repository. https://ncbi.github.io/magicblast
2. Samtools and bcftools. http://www.htslib.org/
3. NCBI Graphical Sequence Viewer. ftp://ftp.ncbi.nlm.nih.gov/pub/factsheets/Factsheet_Graphical_SV.pdf
4. NCBI Genome Data Viewer. ftp://ftp.ncbi.nlm.nih.gov/pub/factsheets/Factsheet_GenomeDataViewer.pdf